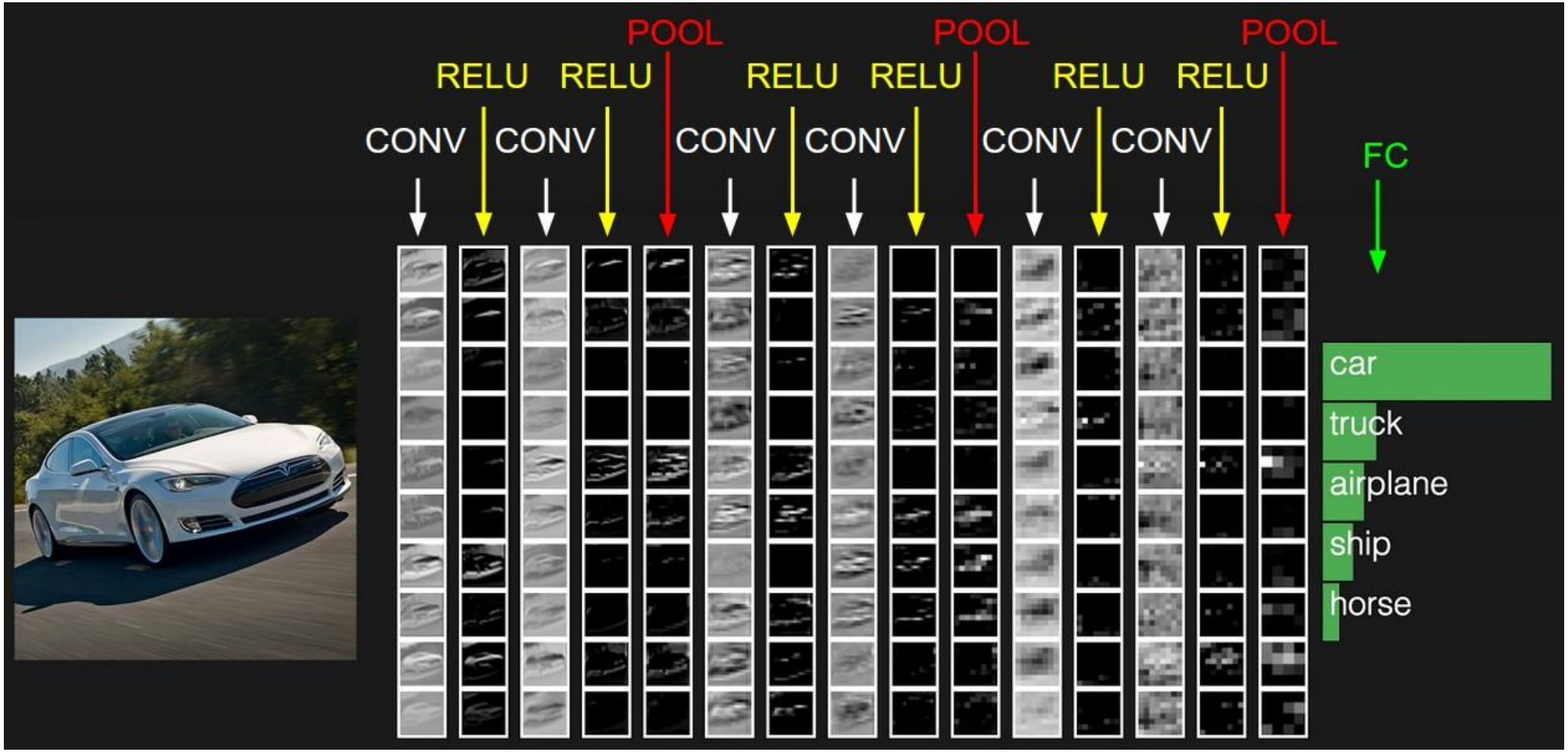
IESL 2017 Final Project
CNN Accelerator

Pham Duc An

Specifications

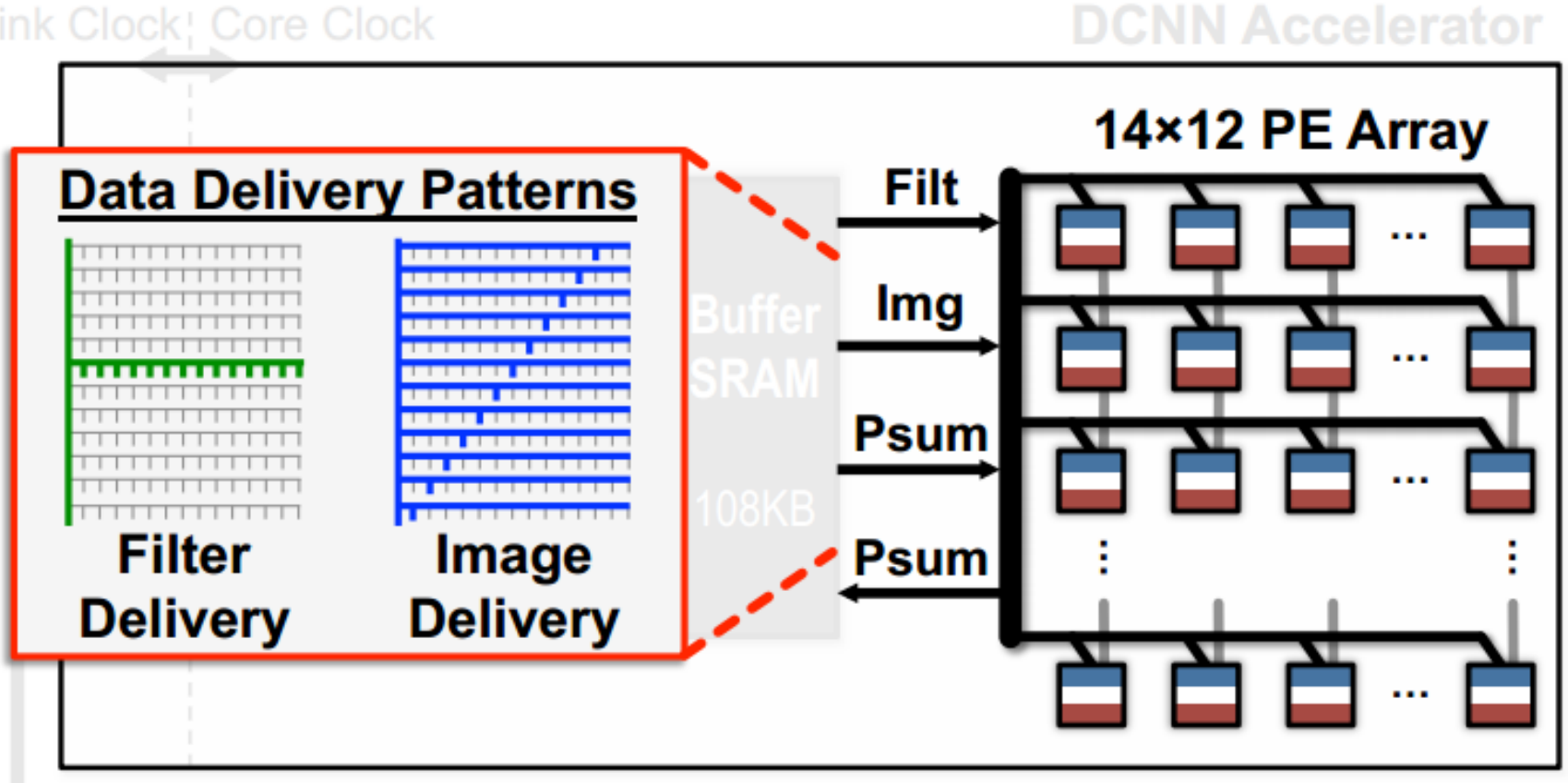
- Implement a CNN accelerator
- SystemC
- Transaction modeling
- TLM 2.0
- Loosely-timed

CNN

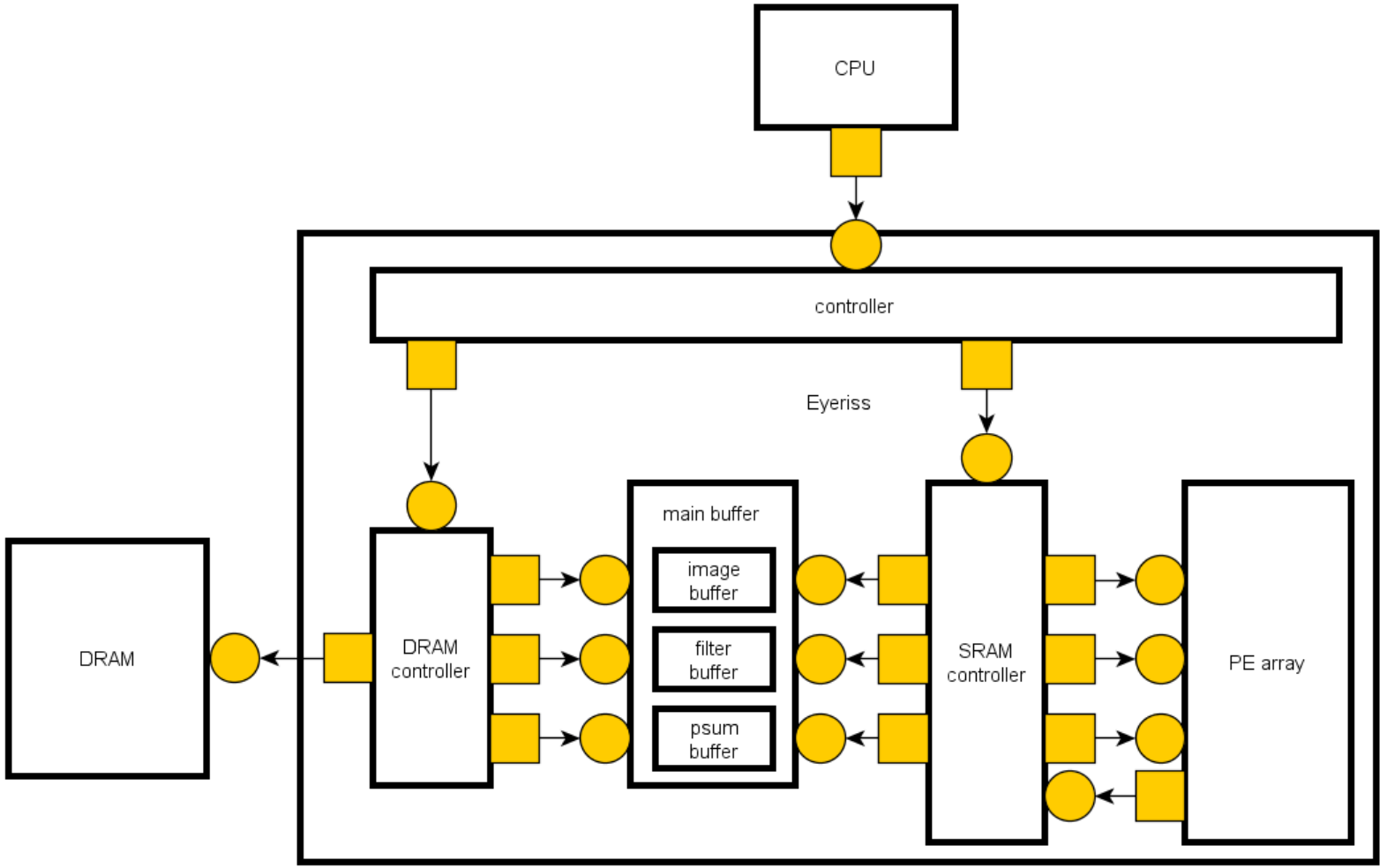

POOL
POOL
POOL
RELU
RELU
RELU
RELU
RELU
RELU
CONV
CONV
CONV
CONV
CONV
CONV
FC
car
truck
airplane
ship
horse

Architecture

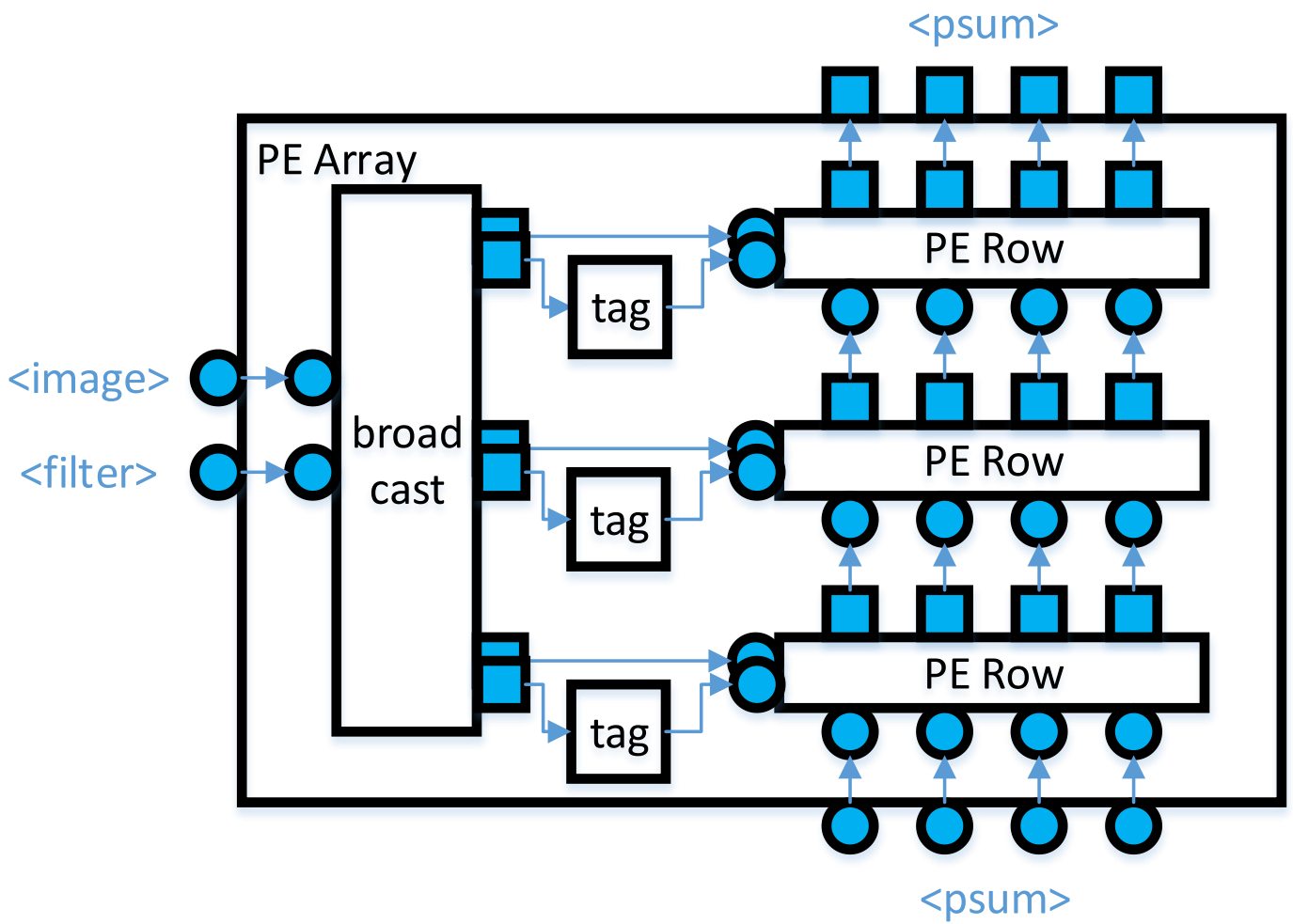
- Based on Eyeriss [1]



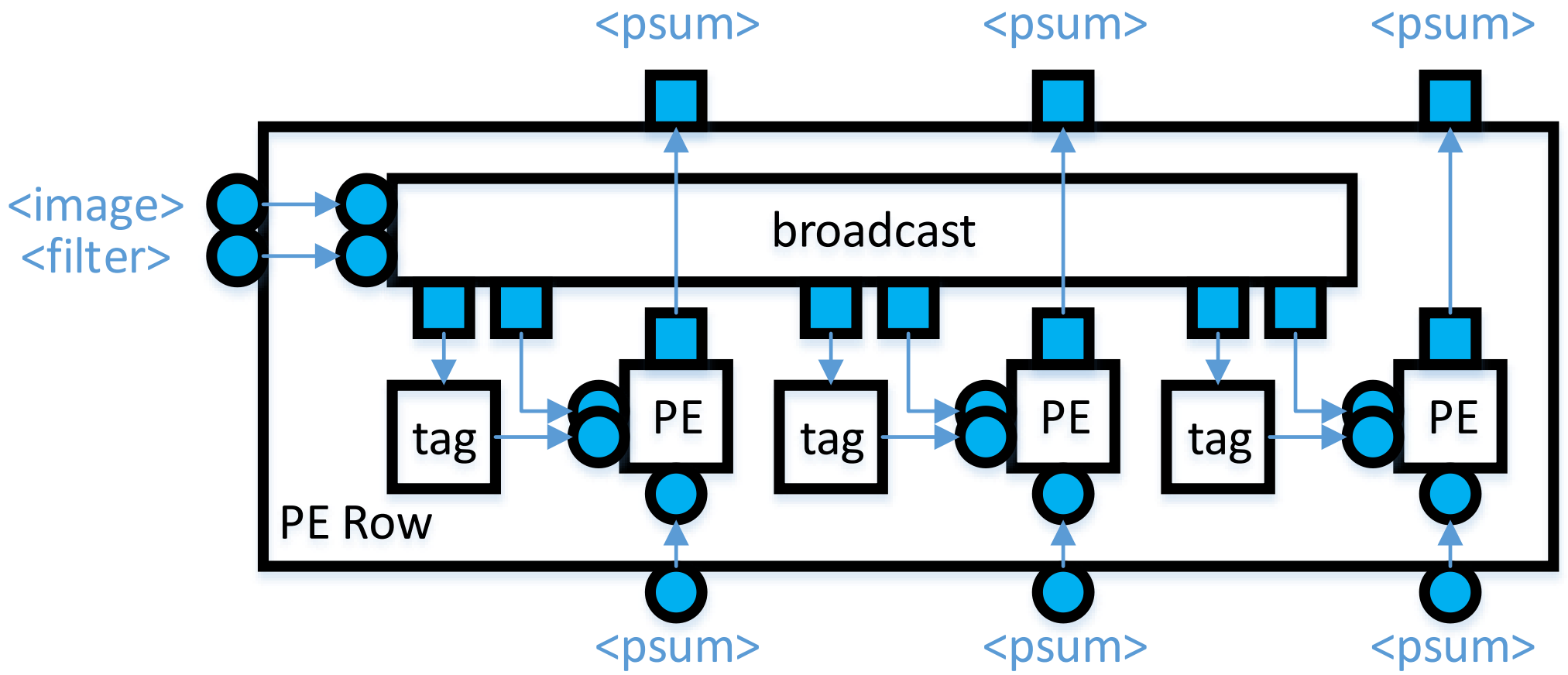
System Model



PE Array


<psum>
PE Array
PE Row
tag
<image>
broad
cast
<filter>
PE Row
tag
PE Row
tag
<psum>

PE Row


<psum>
<psum>
<psum>
<image>
<filter>
broadcast
tag
PE
tag
PE
tag
PE
PE Row
<psum>
<psum>
<psum>

Data Scheme

- Input stationary at buffer level: keep the whole input in the main buffer
- Output stationary at PE level: keep the output result and accumulate it at PE

Testing Result [1/2]

- With 5x5 Eyeriss with only 3x3 PEs are activated

```
27 0 s: Info: top.dram: address=100, data=118.000000
28 0 s: Info: top.dram: Dumping RAM data from address=4096 to address=4135
29 0 s: Info: top.dram: address=4096, data=103.000000
30 0 s: Info: top.dram: address=4100, data=198.000000
31 0 s: Info: top.dram: address=4104, data=105.000000
32 0 s: Info: top.dram: address=4108, data=115.000000
33 0 s: Info: top.dram: address=4112, data=81.000000
34 0 s: Info: top.dram: address=4116, data=255.000000
35 0 s: Info: top.dram: address=4120, data=74.000000
36 0 s: Info: top.dram: address=4124, data=236.000000
37 0 s: Info: top.dram: address=4128, data=41.000000
38 0 s: Info: top.dram: address=4132, data=41.000000
39 16290 ps: Info: /OSCI/SystemC: Simulation stopped by user.
40 16290 ps: Info: top.dram: Dumping RAM data from address=8192 to address=8227
41 16290 ps: Info: top.dram: address=8192, data=220379.000000
42 16290 ps: Info: top.dram: address=8196, data=171397.000000
43 16290 ps: Info: top.dram: address=8200, data=211277.000000
44 16290 ps: Info: top.dram: address=8204, data=205046.000000
45 16290 ps: Info: top.dram: address=8208, data=220331.000000
46 16290 ps: Info: top.dram: address=8212, data=204345.000000
47 16290 ps: Info: top.dram: address=8216, data=212211.000000
48 16290 ps: Info: top.dram: address=8220, data=220743.000000
49 16290 ps: Info: top.dram: address=8224, data=246021.000000
```

Testing Result [1/2]

- AlexNet layer 3

```
55158481920 ps: Info: top.dram: address=0x616AA8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AAC, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AB0, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AB4, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AB8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616ABC, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AC0, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AC4, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AC8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616ACC, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AD0, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AD4, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AD8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616ADC, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AE0, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AE4, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AE8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AEC, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AF0, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AF4, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AF8, data=0.000000
55158481920 ps: Info: top.dram: address=0x616AFC, data=0.000000
55158481920 ps: Info: top.dram: Total accesses: 387456
55158481920 ps: Info: top.dram: Total power consummed: 77491200
55158481920 ps: Info: top.eyeriss.main_buffer: Total accesses: 17464896
55158481920 ps: Info: top.eyeriss.main_buffer: Total power consummed: 104789376
~
```

Evaluation & Conclusion

- The paper itself [1] has some missing information
- The testing result on AlexNet is only with single layer
- Locality is high, the power consumed by on-chip buffer is much higher than off-chip memory

References

[1] Chen, Yu-Hsin, Joel Emer, and Vivienne Sze. "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks." *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. IEEE, 2016.